**National Aeronautics and
Space Administration**

**Jet Propulsion Laboratory**
California Institute of Technology
Pasadena, California

# Analytics Center Framework for Estimating the Circulation and Climate of the Ocean

**Thomas Huang**
thomas.huang@jpl.nasa.gov

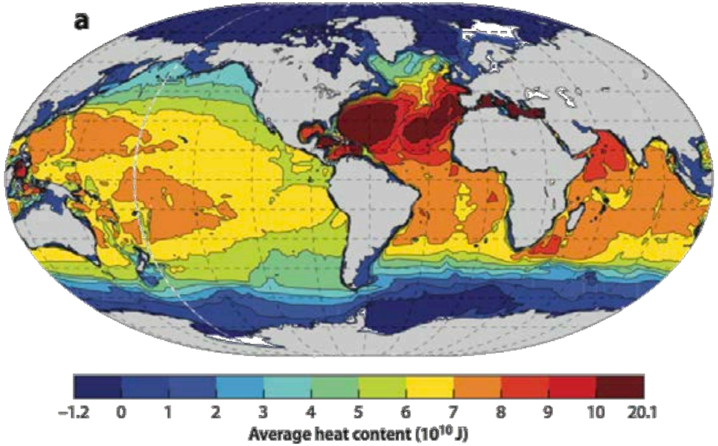
Group Supervisor - Computer Science for Data-Intensive Applications
Strategic Lead - Interactive Data Analytics

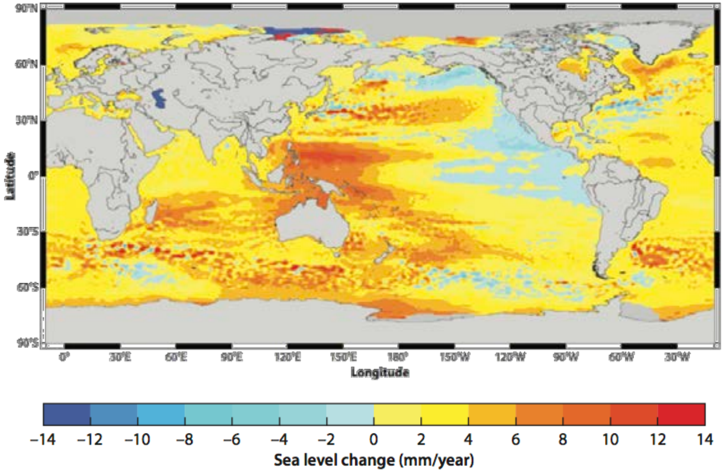
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsements by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

# What is ECCO?

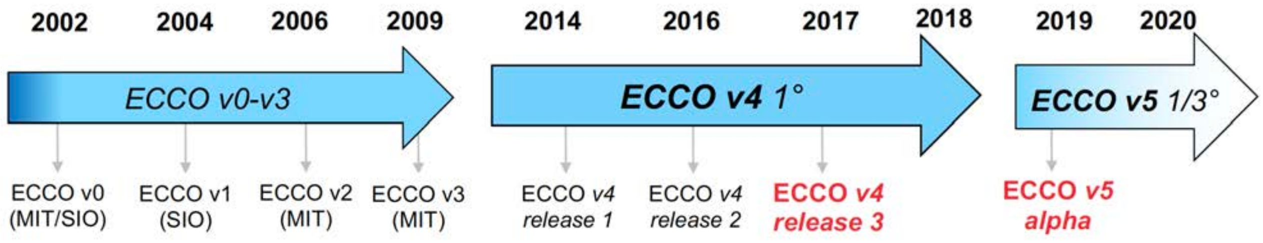
- **Estimating the Circulation and Climate of the Ocean** (ECCO) is a consortium endeavors to produce the best possible estimates of ocean circulation and its role in climate
- Combining state-of-the-art ocean circulation models with global ocean and sea-ice data in a physically and statistically consistent manner
- ECCO products are being used in studies on ocean variability, biological cycles, coastal physics, water cycle, ocean-cryosphere interactions, and geodesy





# ECCO Central Production Timeline

- ECCO v4 is the latest release of ocean state estimate
- It is the first adjoint-based, multi-decadal global ocean and sea-ice state estimate



2002 2004 2006 2009 2014 2016 2017 2018 2019 2020

*ECCO v0-v3*

***ECCO v4** 1°*

***ECCO v5** 1/3°*

ECCO v0 (MIT/SIO)

ECCO v1 (SIO)

ECCO v2 (MIT)

ECCO v3 (MIT)

ECCO v4 *release 1*

ECCO v4 *release 2*

**ECCO v4 *release 3***

**ECCO v5 *alpha***

**ECCO v4r3**

- A 1-deg resolution 4D (space and time) **reconstruction** of the **1992 – 2015 global ocean** and **sea-ice state**
- With over 80 variables

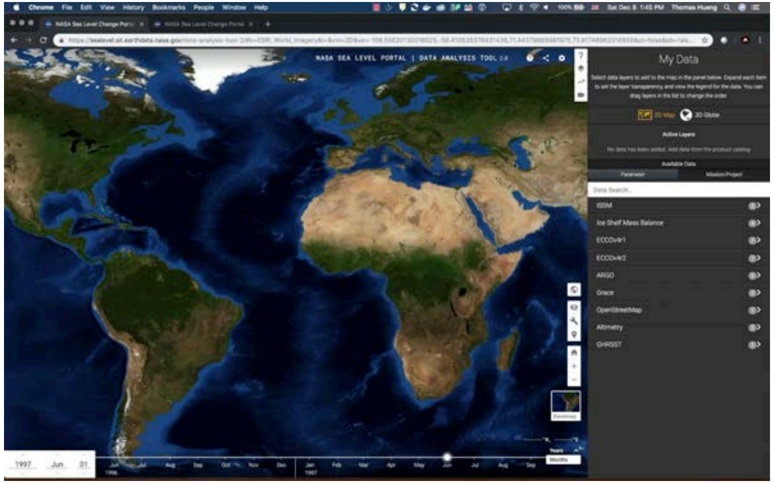
- How to visualize and analyze 80 variables?
- How to compare them?
- How to compare ECCO variables with other observational variables to see how they interact?

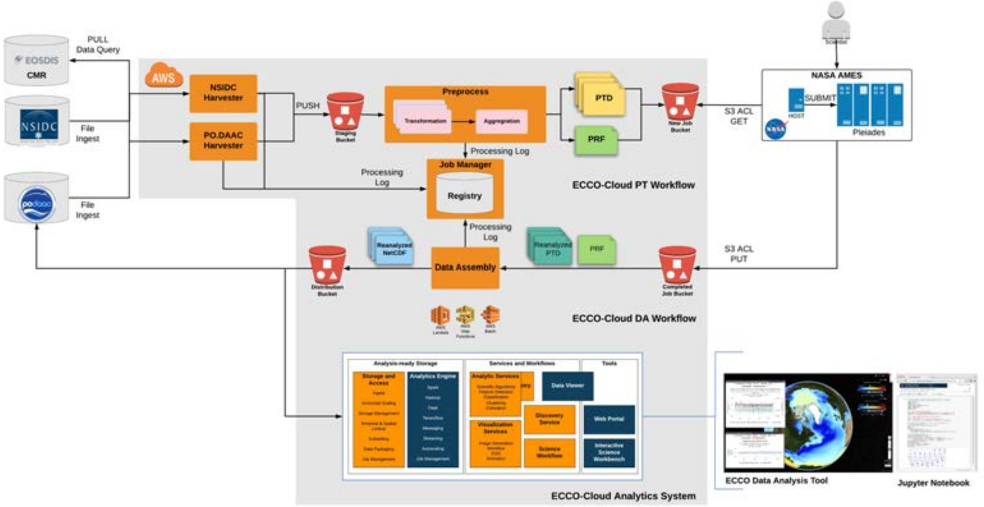
National Aeronautics and
Space Administration

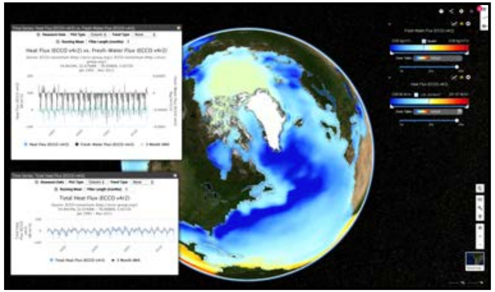
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

# Estimating the Circulation and Climate of the Ocean

NASA ACCESS Program | PI: Patrick Heimbach; Co-Is: Ian Fenty and Thomas Huang

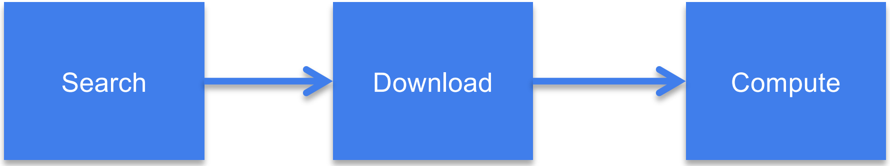




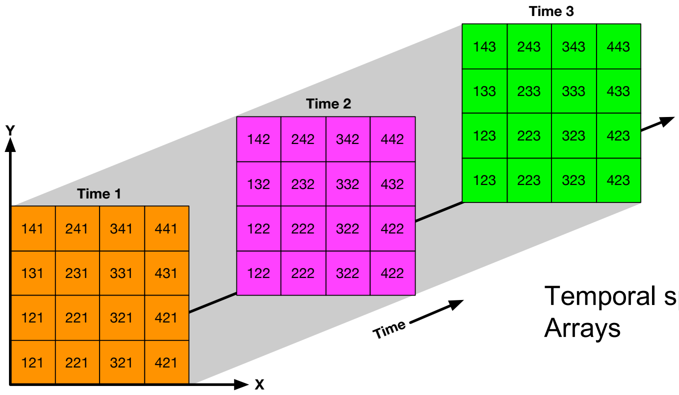


- The **Estimating the Circulation and Climate of the Ocean (ECCO)** global ocean state estimation system (https://ecco.jpl.nasa.gov) is the premier tool for synthesizing NASA's diverse Earth system observations into a complete physical description of Earth's time-evolving full-depth ocean and sea ice system.
- Automate generation of ECCO reanalysis products into CF-compliant netCDF products
- Integrating Amazon Cloud with NASA Ames Pleiades petascale supercomputer
- Establish ECCO Data Analysis Services and web portal for interactive visualization and analysis, and distribution
- Support multi-dimensional data visualization and analysis

# Traditional Method for Analyze Satellite Measurements


Search → Download → Compute

- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files



Time 1

| | | | |
|---|---|---|---|
| 141 | 241 | 341 | 441 |
| 131 | 231 | 331 | 431 |
| 121 | 221 | 321 | 421 |
| 121 | 221 | 321 | 421 |

Time 2

| | | | |
|---|---|---|---|
| 142 | 242 | 342 | 442 |
| 132 | 232 | 332 | 432 |
| 122 | 222 | 322 | 422 |
| 122 | 222 | 322 | 422 |

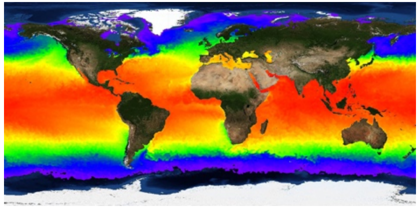
Time 3

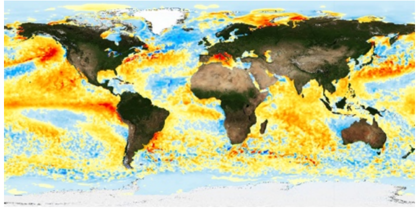
| | | | |
|---|---|---|---|
| 143 | 243 | 343 | 443 |
| 133 | 233 | 333 | 433 |
| 123 | 223 | 323 | 423 |
| 123 | 223 | 323 | 423 |

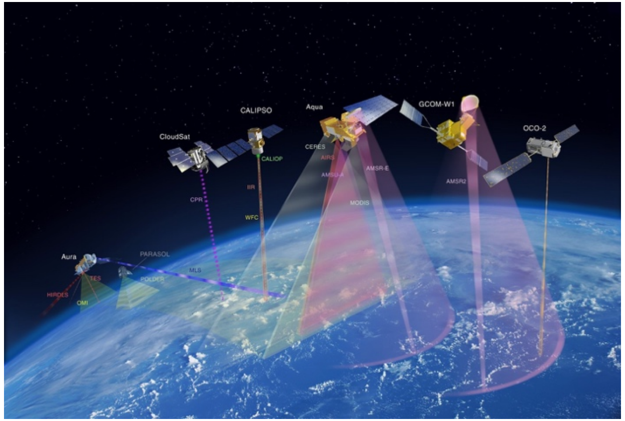
Temporal spatial Arrays

### Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck







National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

# Processors are not Getting Faster

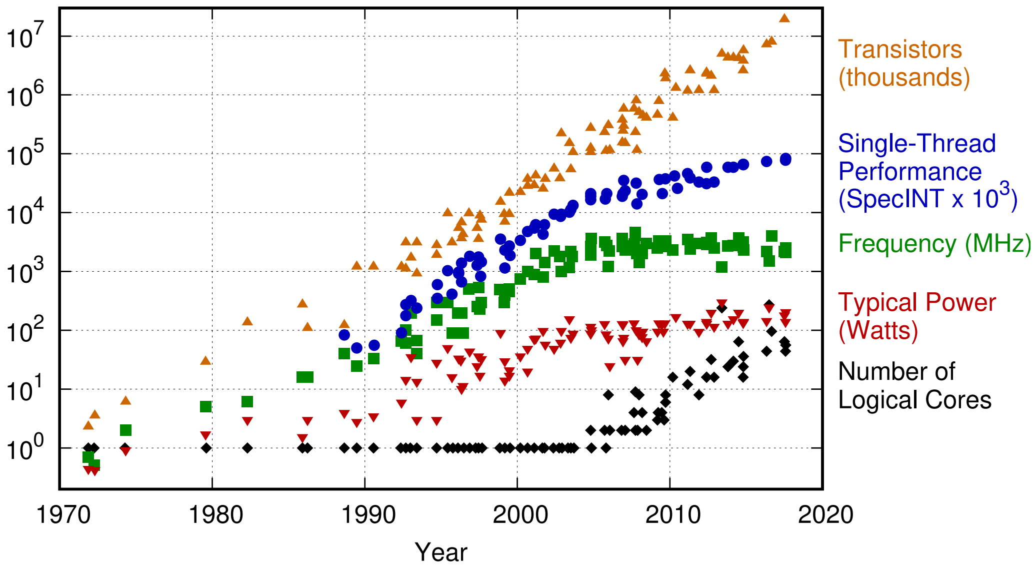
2004: First Pentium 4 processor with 3.0GHz clock speed

2018: Apple’s MacBook Pro has clock speed of 2.7GHz

14 years later, not much has gain in raw processing power

**Modern big data architects are required to “think outside of the box”. Literally!**

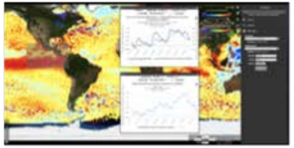
42 Years of Microprocessor Trend Data



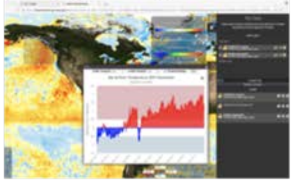
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

# Enabling Next Generation of Earth Science Tools and Services

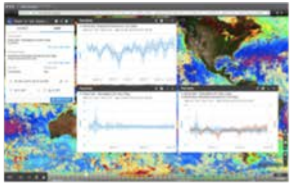
NASA Sea Level Change Portal



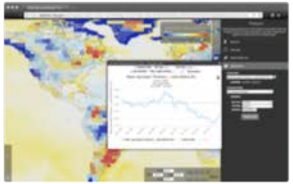
Oceanographic Anomaly Detection



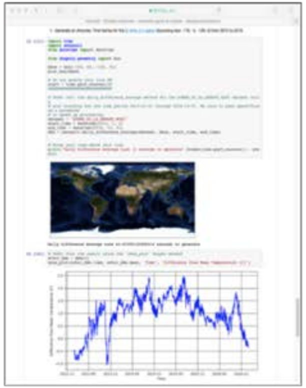
PO.DAAC State Of The Ocean



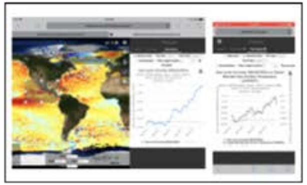
Hydrological Basin Analysis



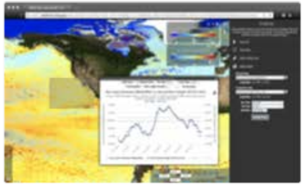
Jupyter Notebook - Interactive Workbench



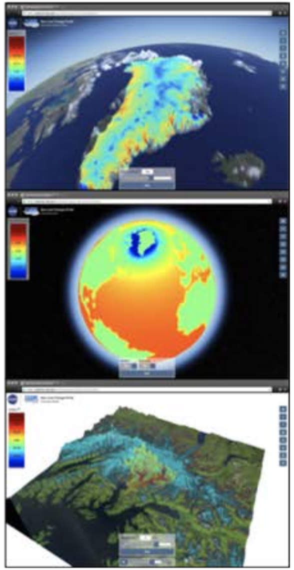
Mobile Analysis



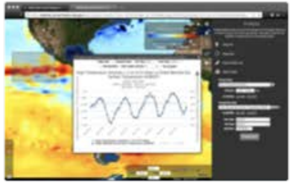
In Situ Data Analysis



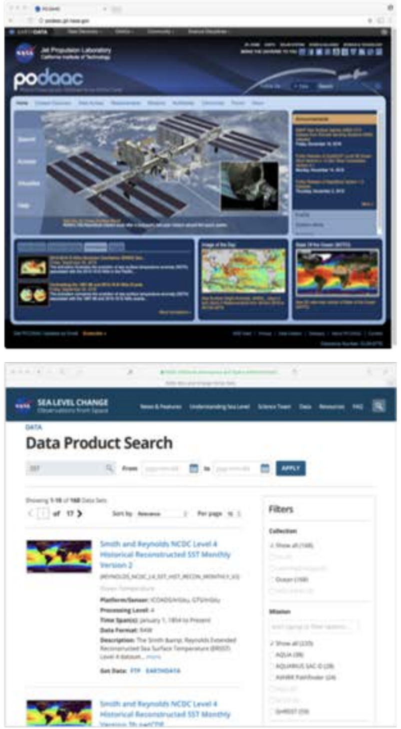
Model Simulations



Model - Observation Comparison



Integrated Search and Discovery



Data Product Search

SST From to APPLY

Showing 1-10 of 168 Data Sets

of 17 Sort by Relevance Per page 10

Smith and Reynolds NCDC Level 4 Historical Reconstructed SST Monthly Version 2

Platform/Sensor: ICOADS/In Situ, GTS/In Situ

Processing Level: 4

Time Span(s): January 1, 1854 to Present

Data Format: RAW

Description: The Smith &amp; Reynolds Extended Reconstructed Sea Surface Temperature (ERSST) Level 4 dataset... more

Get Data: FTP EARTHDATA

Smith and Reynolds NCDC Level 4 Historical Reconstructed SST Monthly

Filters

Collection

✓ Show all (168)

Ocean (168)

Mission

✓ Show all (220)

AQUA (38)

AQUARIUS SAC-D (28)

AVHRR Pathfinder (24)

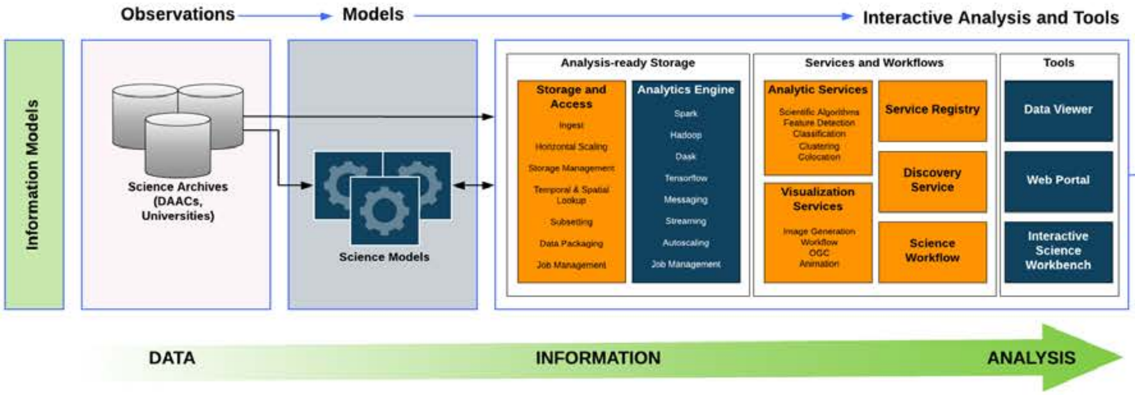
GHRSST (59)

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

# Integrated Science Data Analytics Platform

## Creating SaaS and PaaS for Science Tools and Services



Systematic analysis of massive data

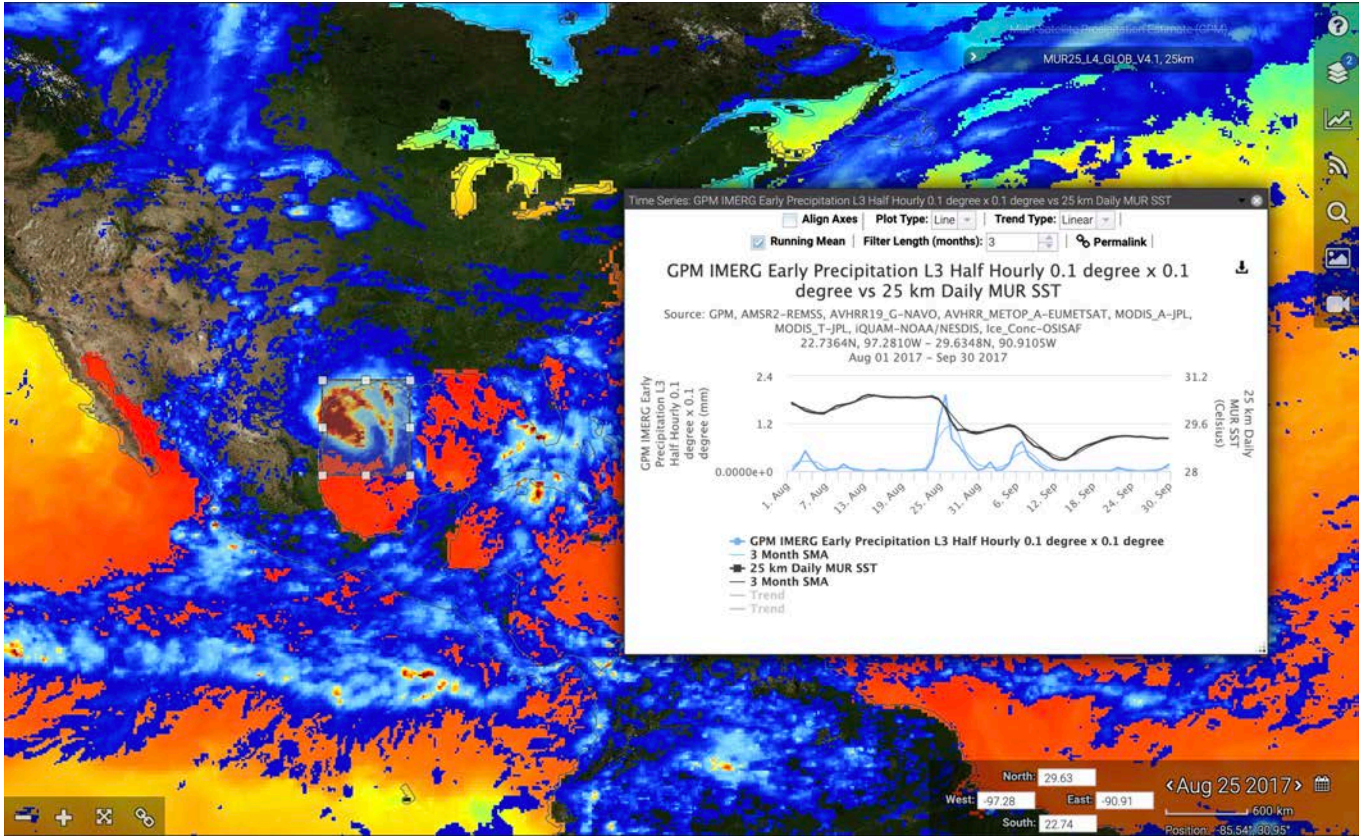
- **Integrated Science Data Analytics Platform**: an analytic center framework to provide an environment for conducting a science investigation
    - Enables the confluence of resources for that investigation
    - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the research community to focus on the investigation
- Scale computational and data infrastructures
- Shift towards integrated data analytics
- Algorithms for identifying and extracting interesting features and patterns

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

# Analyze Hurricane Harvey using GPM and SST

## Aug 17, 2017 – Sept. 2, 2017



Time Series: GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree vs 25 km Daily MUR SST

□ Align Axes | Plot Type: Line | Trend Type: Linear |

☑ Running Mean | Filter Length (months): 3 | Permalink |

GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree vs 25 km Daily MUR SST

Source: GPM, AMSR2-REMSS, AVHRR19_G-NAVO, AVHRR_METOP_A-EUMETSAT, MODIS_A-JPL, MODIS_T-JPL, iQUAM-NOAA/NESDIS, Ice_Conc-OSISAF
22.7364N, 97.2810W - 29.6348N, 90.9105W
Aug 01 2017 - Sep 30 2017

North: 29.63
West: -97.28
East: -90.91
South: 22.74

‹Aug 25 2017›

Position: -85.54°, 30.95°

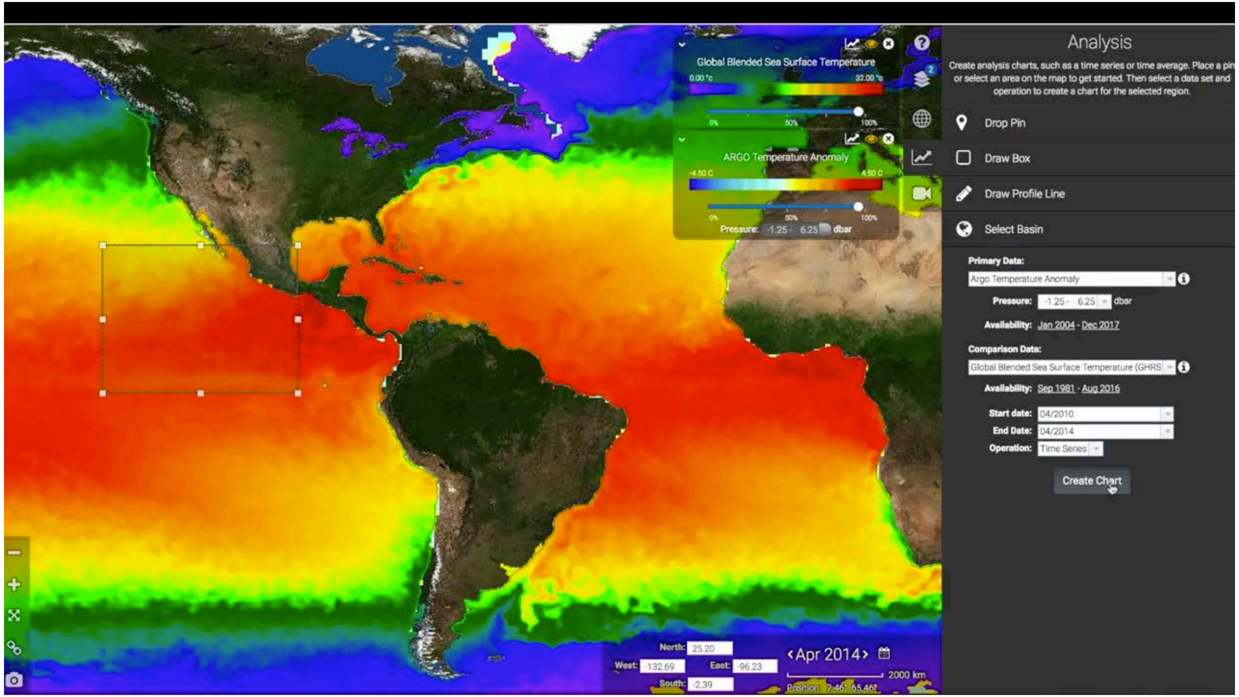
Multi-Satellite Precipitation Estimate (GPM)
MUR25_L4_GLOB_V4.1, 25km
2
GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree (mm)
2.4
1.2
0.0000e+0
25 km Daily MUR SST (Celsius)
31.2
29.6
28
1. Aug
7. Aug
13. Aug
19. Aug
25. Aug
31. Aug
6. Sep
12. Sep
18. Sep
24. Sep
30. Sep
GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree
3 Month SMA
25 km Daily MUR SST
3 Month SMA
Trend
Trend
600 km

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

# Visualize and Analyze Sea Level

## NASA Sea Level Change Portal - https://sealevel.nasa.gov



Pressure: -1.25 - 6.25 dbar

North: 25.20
West: -132.69
East: -96.23
South: -2.39

Analysis

Create analysis charts, such as a time series or time average. Place a pin or select an area on the map to get started. Then select a data set and operation to create a chart for the selected region.

- Drop Pin
- Draw Box
- Draw Profile Line
- Select Basin

**Primary Data:** Argo Temperature Anomaly

**Pressure:** -1.25 - 6.25 dbar

**Availability:** Jan 2004 - Dec 2017

**Comparison Data:** Global Blended Sea Surface Temperature (GHRS

**Availability:** Sep 1981 - Aug 2016

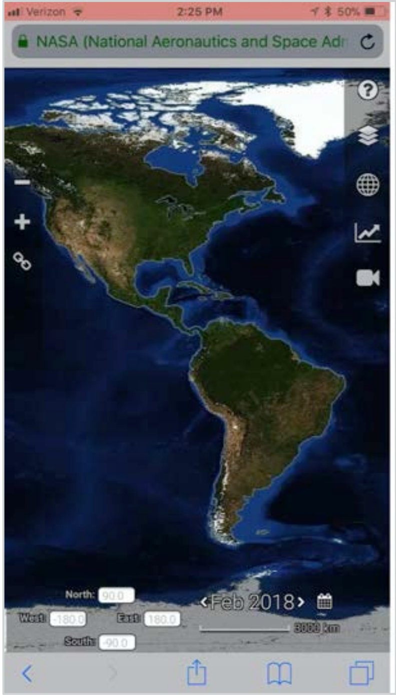
**Start date:** 04/2010

**End Date:** 04/2014

**Operation:** Time Series

Create Chart

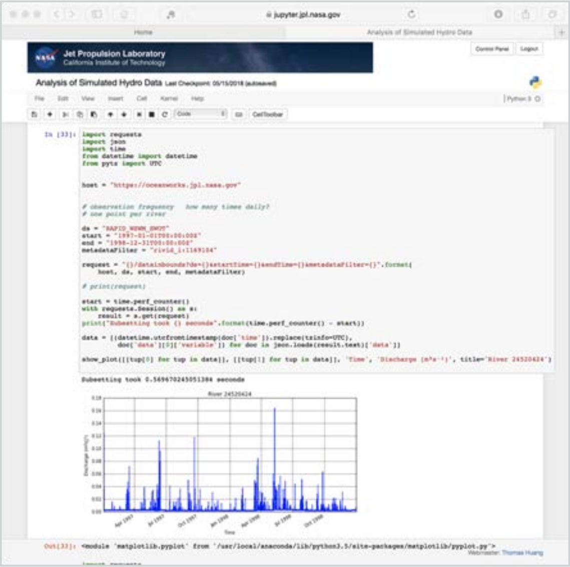
Analyze *in situ* and satellite observations



North: 90.0
West: -180.0
East: 180.0
South: -90.0

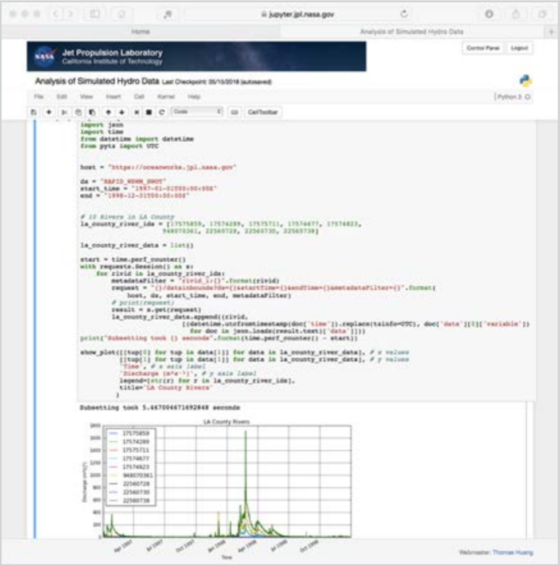
Analyze Sea Level
on mobiles

# Hydrology


jupyter.jpl.nasa.gov
Home
Analysis of Simulated Hydro Data
Jet Propulsion Laboratory
California Institute of Technology
Control Panel
Logout
Analysis of Simulated Hydro Data Last Checkpoint: 05/15/2018 (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Code
CellToolbar
River 24520424
Time
Webmaster: Thomas Huang

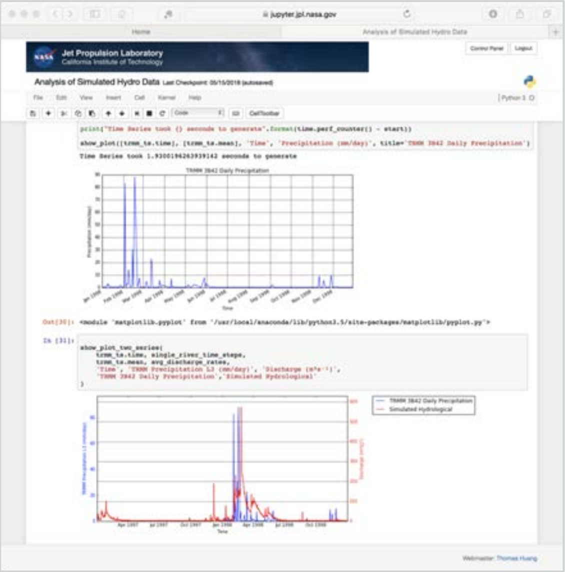
```
In [33]: import requests
         import json
         import time
         from datetime import datetime
         from pytz import UTC

         host = "https://oceanworks.jpl.nasa.gov"

         # observation frequency   how many times daily?
         # one point per river

         ds = "RAPID_WSWM_SWOT"
         start = "1997-01-01T00:00:00Z"
         end = "1998-12-31T00:00:00Z"
         metadataFilter = "rivid_i:1169104"

         request = "{}/datainbounds?ds={}&startTime={}&endTime={}&metadataFilter={}".format(
             host, ds, start, end, metadataFilter)

         # print(request)

         start = time.perf_counter()
         with requests.Session() as s:
             result = s.get(request)
         print("Subsetting took {} seconds".format(time.perf_counter() - start))

         data = [(datetime.utcfromtimestamp(doc['time']).replace(tzinfo=UTC),
                   doc['data'][0]['variable']) for doc in json.loads(result.text)['data']]

         show_plot([[tup[0] for tup in data]], [[tup[1] for tup in data]], 'Time', 'Discharge (m³s⁻¹)', title='River 24520424')

         Subsetting took 0.569670245051384 seconds

Out[33]: <module 'matplotlib.pyplot' from '/usr/local/anaconda/lib/python3.5/site-packages/matplotlib/pyplot.py'>
```

**Retrieval of a single river time series**


jupyter.jpl.nasa.gov
Home
Analysis of Simulated Hydro Data
Jet Propulsion Laboratory
California Institute of Technology
Control Panel
Logout
Analysis of Simulated Hydro Data Last Checkpoint: 05/15/2018 (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Code
CellToolbar
LA County Rivers
17575859
17574289
17575711
17574677
17574823
948070361
22560728
22560730
22560738
Time
Webmaster: Thomas Huang

```
import json
import time
from datetime import datetime
from pytz import UTC

host = "https://oceanworks.jpl.nasa.gov"

ds = "RAPID_WSWM_SWOT"
start_time = "1997-01-01T00:00:00Z"
end = "1998-12-31T00:00:00Z"

# 10 Rivers in LA County
la_county_river_ids = [17575859, 17574289, 17575711, 17574677, 17574823,
                       948070361, 22560728, 22560730, 22560738]

la_county_river_data = list()

start = time.perf_counter()
with requests.Session() as s:
    for rivid in la_county_river_ids:
        metadataFilter = "rivid_i:{}".format(rivid)
        request = "{}/datainbounds?ds={}&startTime={}&endTime={}&metadataFilter={}".format(
            host, ds, start_time, end, metadataFilter)
        # print(request)
        result = s.get(request)
        la_county_river_data.append((rivid,
                                  [(datetime.utcfromtimestamp(doc['time']).replace(tzinfo=UTC), doc['data'][0]['variable'])
                                   for doc in json.loads(result.text)['data']]))
print("Subsetting took {} seconds".format(time.perf_counter() - start))

show_plot([[tup[0] for tup in data[1]] for data in la_county_river_data], # x values
          [[tup[1] for tup in data[1]] for data in la_county_river_data], # y values
          'Time', # x axis label
          'Discharge (m³s⁻¹)', # y axis label
          legend=[str(r) for r in la_county_river_ids],
          title='LA County Rivers'
         )

Subsetting took 5.467004471492848 seconds
```

**Retrieval of time series from 9 rivers**


jupyter.jpl.nasa.gov
Home
Analysis of Simulated Hydro Data
Jet Propulsion Laboratory
California Institute of Technology
Control Panel
Logout
Analysis of Simulated Hydro Data Last Checkpoint: 05/15/2018 (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Code
CellToolbar
TRMM 3B42 Daily Precipitation
Time
TRMM 3B42 Daily Precipitation
Simulated Hydrological
Webmaster: Thomas Huang

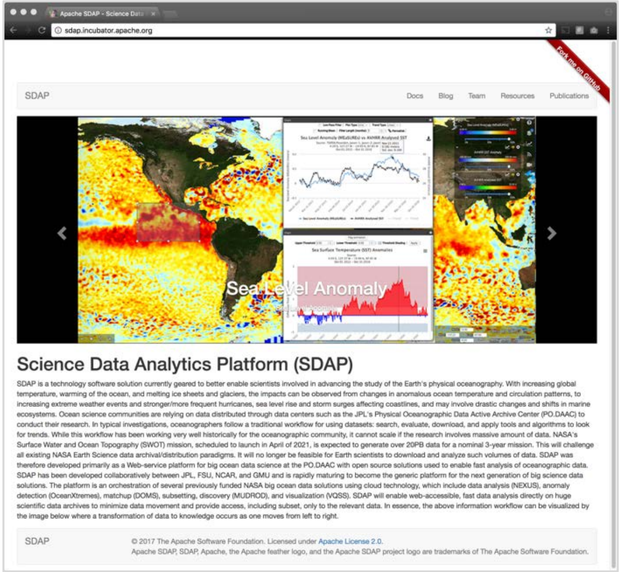
```
print("Time Series took {} seconds to generate".format(time.perf_counter() - start))

show_plot([trmm_ts.time], [trmm_ts.mean], 'Time', 'Precipitation (mm/day)', title='TRMM 3B42 Daily Precipitation')

Time Series took 1.9300196263939142 seconds to generate

Out[30]: <module 'matplotlib.pyplot' from '/usr/local/anaconda/lib/python3.5/site-packages/matplotlib/pyplot.py'>

In [31]:
show_plot_two_series(
    trmm_ts.time, single_river_time_steps,
    trmm_ts.mean, avg_discharge_rates,
    'Time', 'TRMM Precipitation L3 (mm/day)', 'Discharge (m³s⁻¹)',
    'TRMM 3B42 Daily Precipitation','Simulated Hydrological'
)
```

**Time series coordination between TRMM and river**

- Simulated hydrology data in preparation for SWOT hydrology
- **River data**: ~3.6 billion data points. 3-hour sample rate. Consists of measurements from ~600,000 rivers
- **TRMM data**: 17 years, .25deg, 1.5 billion data points
- Sub-second retrieval of river measurements
- On-the-fly computation of time series and generate coordination plot

# Free and Open Source Software (FOSS)

- After more than two years of active development, on October 2017 the **NASA ESOT/AIST OceanWorks** team established Apache Software Foundation and established the **Science Data Analytics Platform (SDAP)** in the **Apache Incubator**
- Technology sharing through Free and Open Source Software (FOSS)
- Why? Further technology evolution that is restricted by projects / missions
- It is more than GitHub
  - Quarterly reporting
  - Reports are open for community review by over 6000 committers
  - SDAP has a group of appointed international mentors
- **SDAP and many of its affiliated projects are now being developed in the open**
  - Support local cluster and cloud computing platform support
  - Fully containerized using Docker and Kubernetes
  - Infrastructure orchestration using Amazon CloudFormation
  - Satellite and model data analysis: time series, correlation map,
  - In situ data analysis and collocation with satellite measurements
  - Fast data subsetting
  - Upload and execute custom parallel analytic algorithms
  - Data services integration architecture
  - OpenSearch and dynamic metadata translation
  - Mining of user interaction and data to enable discovery and recommendations



## Science Data Analytics Platform (SDAP)

SDAP is a technology software solution currently geared to better enable scientists involved in advancing the study of the Earth's physical oceanography. With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, the impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and stronger/more frequent hurricanes, sea level rise and storm surges affecting coastlines, and may involve drastic changes and shifts in marine ecosystems. Ocean science communities are relying on data distributed through data centers such as the JPL's Physical Oceanographic Data Active Archive Center (PO.DAAC) to conduct their research. In typical investigations, oceanographers follow a traditional workflow for using datasets: search, evaluate, download, and apply tools and algorithms to look for trends. While this workflow has been working very well historically for the oceanographic community, it cannot scale if the research involves massive amount of data. NASA's Surface Water and Ocean Topography (SWOT) mission, scheduled to launch in April of 2021, is expected to generate over 20PB data for a nominal 3-year mission. This will challenge all existing NASA Earth Science data archival/distribution paradigms. It will no longer be feasible for Earth scientists to download and analyze such volumes of data. SDAP was therefore developed primarily as a Web-service platform for big ocean data science at the PO.DAAC with open source solutions used to enable fast analysis of oceanographic data. SDAP has been developed collaboratively between JPL, FSU, NCAR, and GMU and is rapidly maturing to become the generic platform for the next generation of big science data solutions. The platform is an orchestration of several previously funded NASA big ocean data solutions using cloud technology, which include data analysis (NEXUS), anomaly detection (OceanXtremes), matchup (DOMS), subsetting, discovery (MUDROD), and visualization (VQSS). SDAP will enable web-accessible, fast data analysis directly on huge scientific data archives to minimize data movement and provide access, including subset, only to the relevant data. In essence, the above information workflow can be visualized by the image below where a transformation of data to knowledge occurs as one moves from left to right.

SDAP

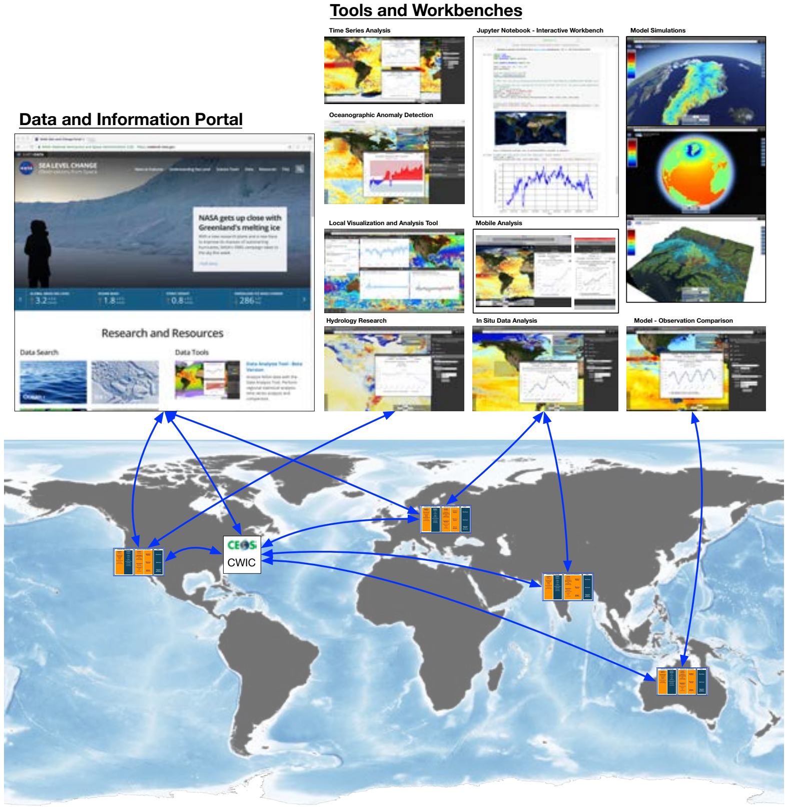
http://sdap.apache.org

APACHE INCUBATOR

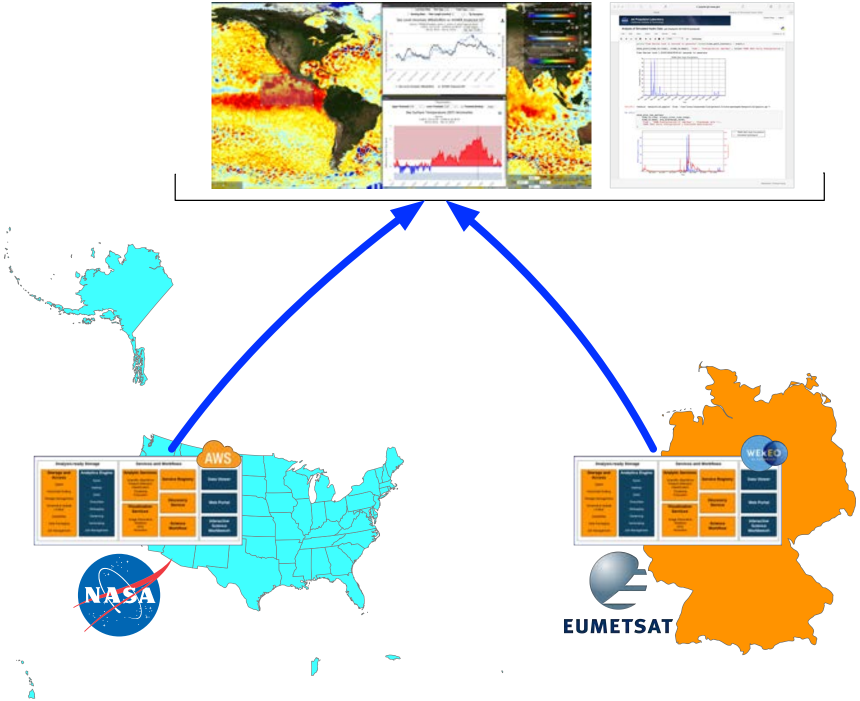
# Other Applications of SDAP

# COVERAGE’s Distributed Analytics Center Architecture

- **Committee of Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEO (COVERAGE) Initiative**
- Seeks to provide **improved access** to **multi-agency ocean remote sensing data** that are **better integrated with in-situ and biological observations**, in support of **oceanographic and decision support applications** for societal benefit.
- A community-support open specification with common taxonomies, information model, and API (maybe security)
- Putting value-added services next to the data to eliminate unnecessary data movement
- Avoid data replication. Reduce unnecessary data movement and egress charges
- Public accessible RESTful analytic APIs where computation is next to the data
- Analytic engine infused and managed by the data centers perhaps on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files


Tools and Workbenches
Time Series Analysis
Jupyter Notebook - Interactive Workbench
Model Simulations
Oceanographic Anomaly Detection
Local Visualization and Analysis Tool
Mobile Analysis
Hydrology Research
In Situ Data Analysis
Model - Observation Comparison
Data and Information Portal
SEA LEVEL CHANGE
NASA gets up close with Greenland's melting ice
Research and Resources
Data Search
Data Tools
CEOS
CWIC

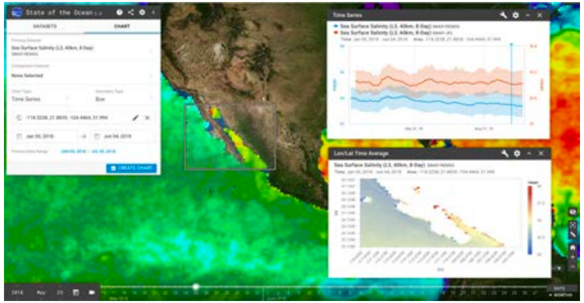
# COVERAGE Phase B

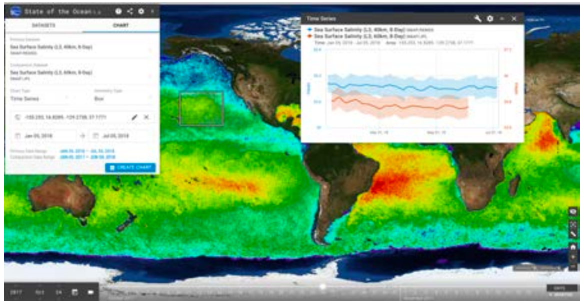


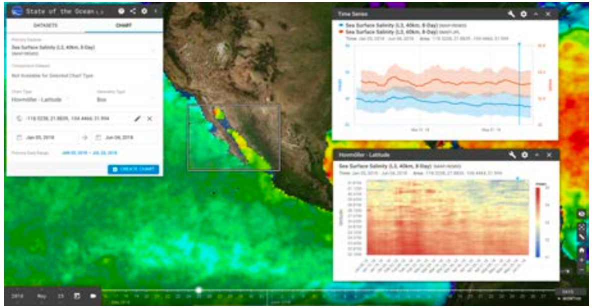
- WEkEO
  - Copernicus Data and Information Access Services (DIAS)
    1. Copernicus Data
    2. Virtual Environment and Tools
    3. User Support
  - Harmonized Data Access for Satellite data and Services
  - Virtualized infrastructure for personal sandboxes
  - Pre-configured tools
- COVERAGE Phase B
  - Establish US Node on Amazon Cloud
  - Establish EU Node on WEkEO at EUMETSAT
  - Establish COVERAGE data portal and analysis tool powered by the COVERAGE Nodes at US and EU

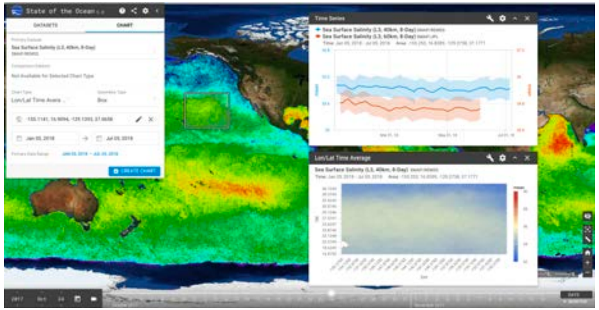
# PO.DAAC’s SOTO

- NASA’s Physical Oceanography Distributed Active Archive Center (PO.DAAC) is an element of the Earth Observing System Data and Information System (EOSDIS)
- PO.DAAC’s mission is to preserve NASA’s ocean and climate data and make these universally accessible and meaningful
- State of the Ocean (SOTO) is a PO.DAAC’s popular visualization tool for the physical oceanography community
- SOTO v5 will be integrated with Apache SDAP and operate on the Amazon Cloud for on-the-fly data analytics









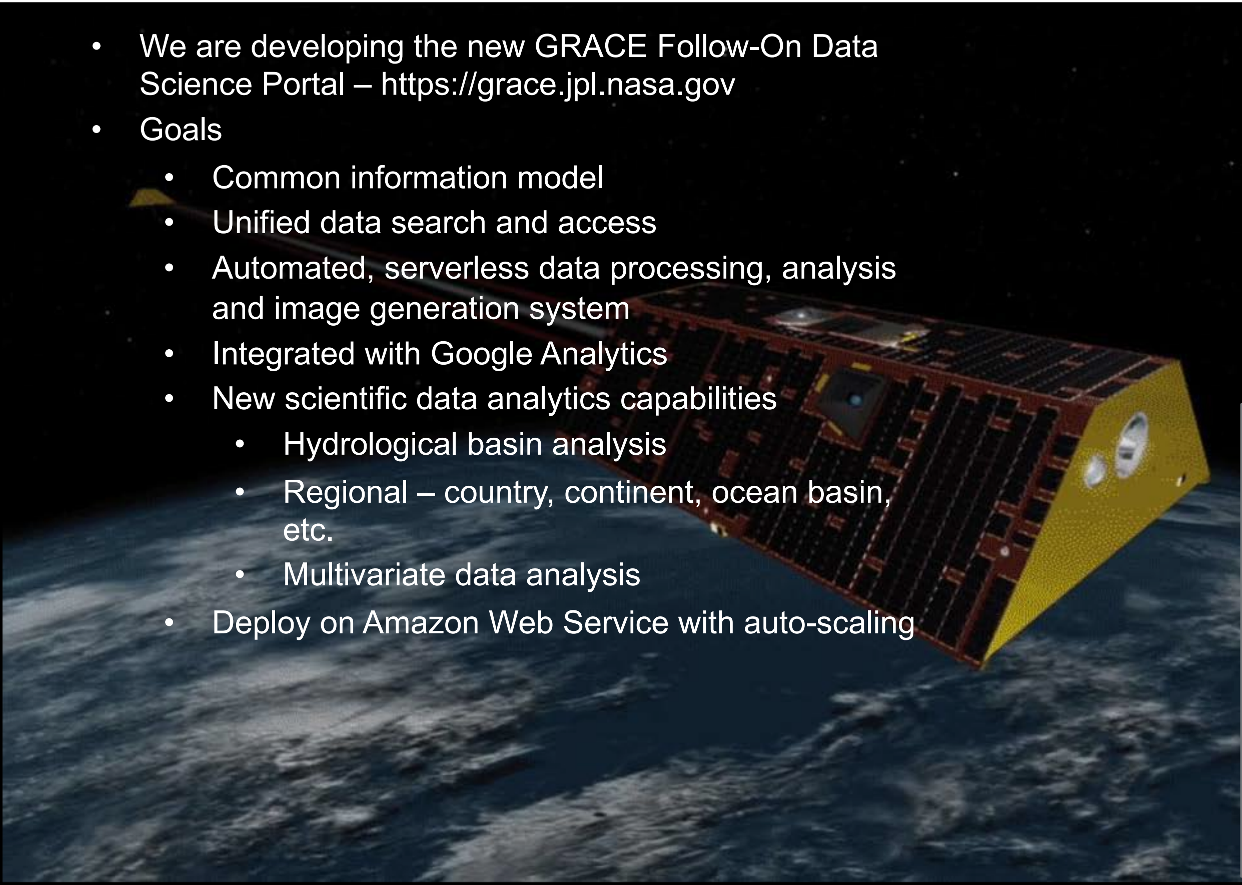
https://podaac.jpl.nasa.gov

National Aeronautics and Space Administration

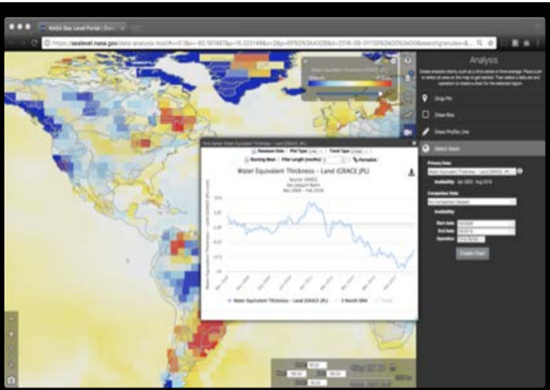
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

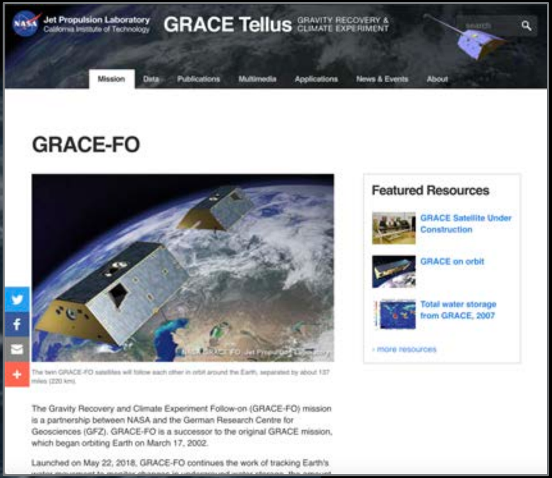
# GRACE-FO Portal and Data Analysis Tool

## GRACE/GRACE-FO Science



- We are developing the new GRACE Follow-On Data Science Portal – https://grace.jpl.nasa.gov
- Goals
  - Common information model
  - Unified data search and access
  - Automated, serverless data processing, analysis and image generation system
  - Integrated with Google Analytics
  - New scientific data analytics capabilities
    - Hydrological basin analysis
    - Regional – country, continent, ocean basin, etc.
    - Multivariate data analysis
  - Deploy on Amazon Web Service with auto-scaling




Jet Propulsion Laboratory
California Institute of Technology
GRACE Tellus
GRAVITY RECOVERY & CLIMATE EXPERIMENT
search
Mission
Data
Publications
Multimedia
Applications
News & Events
About

GRACE-FO

The twin GRACE-FO satellites will follow each other in orbit around the Earth, separated by about 137 miles (220 km).

The Gravity Recovery and Climate Experiment Follow-on (GRACE-FO) mission is a partnership between NASA and the German Research Centre for Geosciences (GFZ). GRACE-FO is a successor to the original GRACE mission, which began orbiting Earth on March 17, 2002.

Launched on May 22, 2018, GRACE-FO continues the work of tracking Earth's

Featured Resources

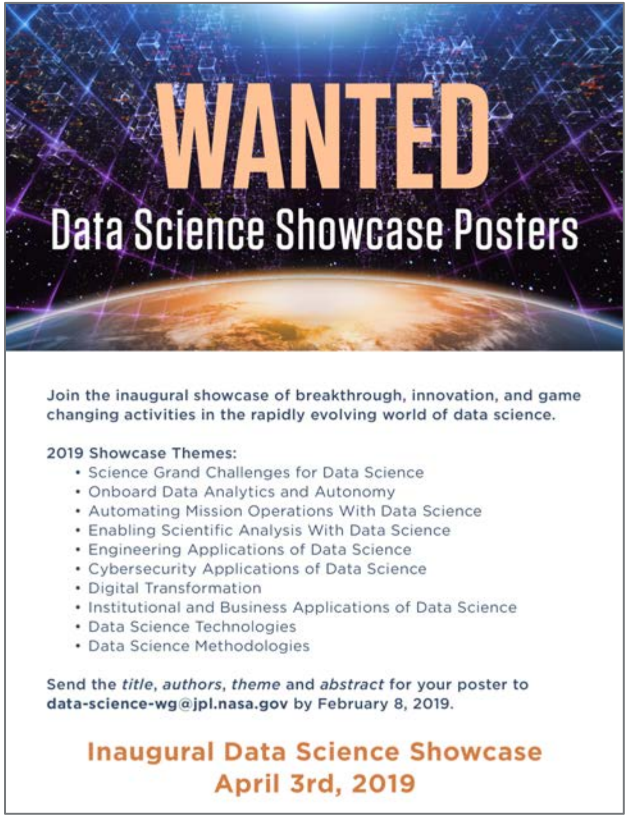
- GRACE Satellite Under Construction
- GRACE on orbit
- Total water storage from GRACE, 2007
- › more resources

# Building Community-Driven Open Data and Open Source Solutions

- Deliver solutions to establish coherent platform solutions
- Embrace open source software
- Community validation
- Evolve the technology through community contributions
- Share recipes and lessons learned
- Technology demonstrations
- Host webinars, hands-on cloud analytics workshops and hackathons

2019 EGU – NASA Hyperwall

Big Data Analytics and Cloud Computing Workshop, 2017 ESIP Summer Meeting, Bloomington, IN



WANTED

Data Science Showcase Posters

Join the inaugural showcase of breakthrough, innovation, and game changing activities in the rapidly evolving world of data science.

2019 Showcase Themes:

- Science Grand Challenges for Data Science
- Onboard Data Analytics and Autonomy
- Automating Mission Operations With Data Science
- Enabling Scientific Analysis With Data Science
- Engineering Applications of Data Science
- Cybersecurity Applications of Data Science
- Digital Transformation
- Institutional and Business Applications of Data Science
- Data Science Technologies
- Data Science Methodologies

Send the *title*, *authors*, *theme* and *abstract* for your poster to data-science-wg@jpl.nasa.gov by February 8, 2019.

Inaugural Data Science Showcase
April 3rd, 2019

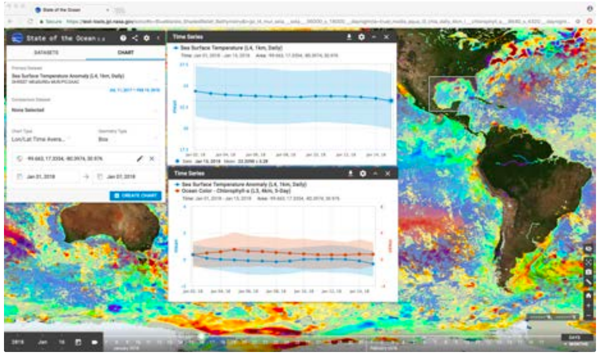
2019 JPL Data Science Showcase

# Partner with NASA and non-NASA Projects - Deliver to Production

- **The gap between visionary to pragmatists is significant.** – Geoffrey Moore
- Become an expert in the production environment and devote resources in automations
- Give project engineering team early access to the PaaS
- Deliver all technical documents and work with project system engineering
- Provide project-focused trainings

NASA's Sea Level Change Team

CEOS SIT Technical Workshop


State of the Ocean
DATASETS
CHART
Sea Surface Temperature Anomaly (L4, 1km, Daily)
None Selected
CREATE CHART
Time Series
Sea Surface Temperature (L4, 1km, Daily)
Time Series
Sea Surface Temperature Anomaly (L4, 1km, Daily)
Ocean Color - Chlorophyll-a (L3, 4km, 5-Day)

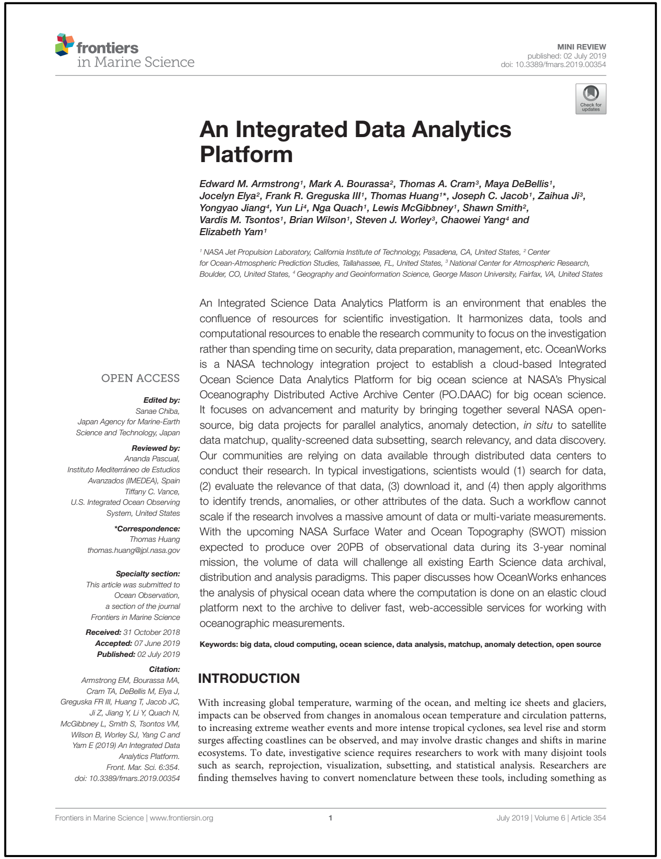
NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)

# OceanObs’19

- Our new Community White Paper: “An Integrated Data Analytics Platform”
- https://www.frontiersin.org/articles/10.3389/fmars.2019.00354/full?&utm_source=Email_to_authors_&utm_medium=Email&utm_content=T1_11.5e1_author&utm_campaign=Email_publication&field=&journalName=Frontiers_in_Marine_Science&id=433796
- We are invited to discuss about “Big Ocean Science Analytics using Apache Science Data Analytics Platform” at OceanObs 2019

OCEAN
OBS’19

AN OCEAN OF OPPORTUNITY
September 16-20, 2019



frontiers
in Marine Science

MINI REVIEW
published: 02 July 2019
doi: 10.3389/fmars.2019.00354

Check for updates

## An Integrated Data Analytics Platform

*Edward M. Armstrong*[1], *Mark A. Bourassa*[2], *Thomas A. Cram*[3], *Maya DeBellis*[1], *Jocelyn Elya*[2], *Frank R. Greguska III*[1], *Thomas Huang*[1]*, *Joseph C. Jacob*[1], *Zaihua Ji*[3], *Yongyao Jiang*[4], *Yun Li*[4], *Nga Quach*[1], *Lewis McGibbney*[1], *Shawn Smith*[2], *Vardis M. Tsontos*[1], *Brian Wilson*[1], *Steven J. Worley*[3], *Chaowei Yang*[4] and *Elizabeth Yam*[1]

*[1] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States, [2] Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, United States, [3] National Center for Atmospheric Research, Boulder, CO, United States, [4] Geography and Geoinformation Science, George Mason University, Fairfax, VA, United States*

OPEN ACCESS

***Edited by:***
*Sanae Chiba,*
*Japan Agency for Marine-Earth Science and Technology, Japan*

***Reviewed by:***
*Ananda Pascual,*
*Instituto Mediterráneo de Estudios Avanzados (IMEDEA), Spain*
*Tiffany C. Vance,*
*U.S. Integrated Ocean Observing System, United States*

****Correspondence:***
*Thomas Huang*
*thomas.huang@jpl.nasa.gov*

***Specialty section:***
*This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science*

***Received:*** *31 October 2018*
***Accepted:*** *07 June 2019*
***Published:*** *02 July 2019*

***Citation:***
*Armstrong EM, Bourassa MA, Cram TA, DeBellis M, Elya J, Greguska FR III, Huang T, Jacob JC, Ji Z, Jiang Y, Li Y, Quach N, McGibbney L, Smith S, Tsontos VM, Wilson B, Worley SJ, Yang C and Yam E (2019) An Integrated Data Analytics Platform. Front. Mar. Sci. 6:354. doi: 10.3389/fmars.2019.00354*

An Integrated Science Data Analytics Platform is an environment that enables the confluence of resources for scientific investigation. It harmonizes data, tools and computational resources to enable the research community to focus on the investigation rather than spending time on security, data preparation, management, etc. OceanWorks is a NASA technology integration project to establish a cloud-based Integrated Ocean Science Data Analytics Platform for big ocean science at NASA’s Physical Oceanography Distributed Active Archive Center (PO.DAAC) for big ocean science. It focuses on advancement and maturity by bringing together several NASA open-source, big data projects for parallel analytics, anomaly detection, *in situ* to satellite data matchup, quality-screened data subsetting, search relevancy, and data discovery. Our communities are relying on data available through distributed data centers to conduct their research. In typical investigations, scientists would (1) search for data, (2) evaluate the relevance of that data, (3) download it, and (4) then apply algorithms to identify trends, anomalies, or other attributes of the data. Such a workflow cannot scale if the research involves a massive amount of data or multi-variate measurements. With the upcoming NASA Surface Water and Ocean Topography (SWOT) mission expected to produce over 20PB of observational data during its 3-year nominal mission, the volume of data will challenge all existing Earth Science data archival, distribution and analysis paradigms. This paper discusses how OceanWorks enhances the analysis of physical ocean data where the computation is done on an elastic cloud platform next to the archive to deliver fast, web-accessible services for working with oceanographic measurements.

**Keywords: big data, cloud computing, ocean science, data analysis, matchup, anomaly detection, open source**

### INTRODUCTION

With increasing global temperature, warming of the ocean, and melting ice sheets and glaciers, impacts can be observed from changes in anomalous ocean temperature and circulation patterns, to increasing extreme weather events and more intense tropical cyclones, sea level rise and storm surges affecting coastlines can be observed, and may involve drastic changes and shifts in marine ecosystems. To date, investigative science requires researchers to work with many disjoint tools such as search, reprojection, visualization, subsetting, and statistical analysis. Researchers are finding themselves having to convert nomenclature between these tools, including something as

Frontiers in Marine Science | www.frontiersin.org 1 July 2019 | Volume 6 | Article 354

# ApacheCon North America

- Invited to discuss our Apache Science Data Analytics Platform (SDAP) project at the ApacheCon North America

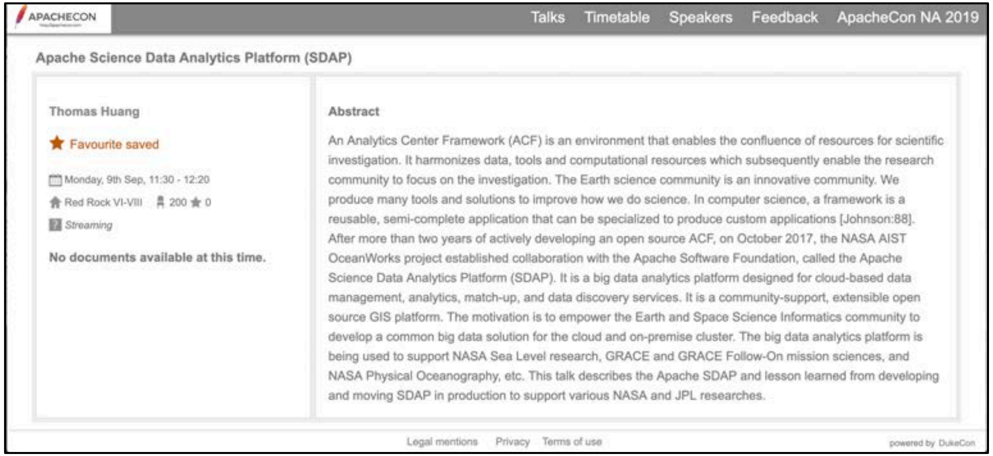
9-12 SEPTEMBER 2019

**ApacheCon North America**

*Las Vegas, NV, USA*

Program announced!

Categories include Big Data, Community, Data Science, Diversity, Geospatial, Graphs, Integration, IoT, Machine Learning, Observability, and much more.



**Apache Science Data Analytics Platform (SDAP)**

**Thomas Huang**

★ Favourite saved

Monday, 9th Sep, 11:30 - 12:20

Red Rock VI-VIII 200 ★ 0

*Streaming*

**No documents available at this time.**

**Abstract**

An Analytics Center Framework (ACF) is an environment that enables the confluence of resources for scientific investigation. It harmonizes data, tools and computational resources which subsequently enable the research community to focus on the investigation. The Earth science community is an innovative community. We produce many tools and solutions to improve how we do science. In computer science, a framework is a reusable, semi-complete application that can be specialized to produce custom applications [Johnson:88]. After more than two years of actively developing an open source ACF, on October 2017, the NASA AIST OceanWorks project established collaboration with the Apache Software Foundation, called the Apache Science Data Analytics Platform (SDAP). It is a big data analytics platform designed for cloud-based data management, analytics, match-up, and data discovery services. It is a community-support, extensible open source GIS platform. The motivation is to empower the Earth and Space Science Informatics community to develop a common big data solution for the cloud and on-premise cluster. The big data analytics platform is being used to support NASA Sea Level research, GRACE and GRACE Follow-On mission sciences, and NASA Physical Oceanography, etc. This talk describes the Apache SDAP and lesson learned from developing and moving SDAP in production to support various NASA and JPL researches.

# Acknowledgement

Ed Armstrong/JPL
Jason Barnett/LARC
Andrew Bingham/JPL
Carmen Boening/JPL
Mark Bourassa/FSU
Mike Chin/JPL
Marge Cole/NASA
Tom Cram/NCAR
Dan Crichton/JPL

Maya DeBellis/JPL
Rich Doyle/JPL
Jocelyn Elya/FSU
Ian Fenty/JPL
Kevin Gill/JPL
Frank Greguska/JPL
Patrick Heimbach/UT Austin
Ben Holt/JPL
Thomas Huang/JPL

Joe Jacob/JPL
Zaihua Ji/NCAR
Yongyao Jiang/GMU
Felix Landerer/JPL
Yun Li/GMU
Eric Lindstrom/NASA
Mike Little/NASA
Chris Lynnes/NASA
Lewis McGibbney/JPL

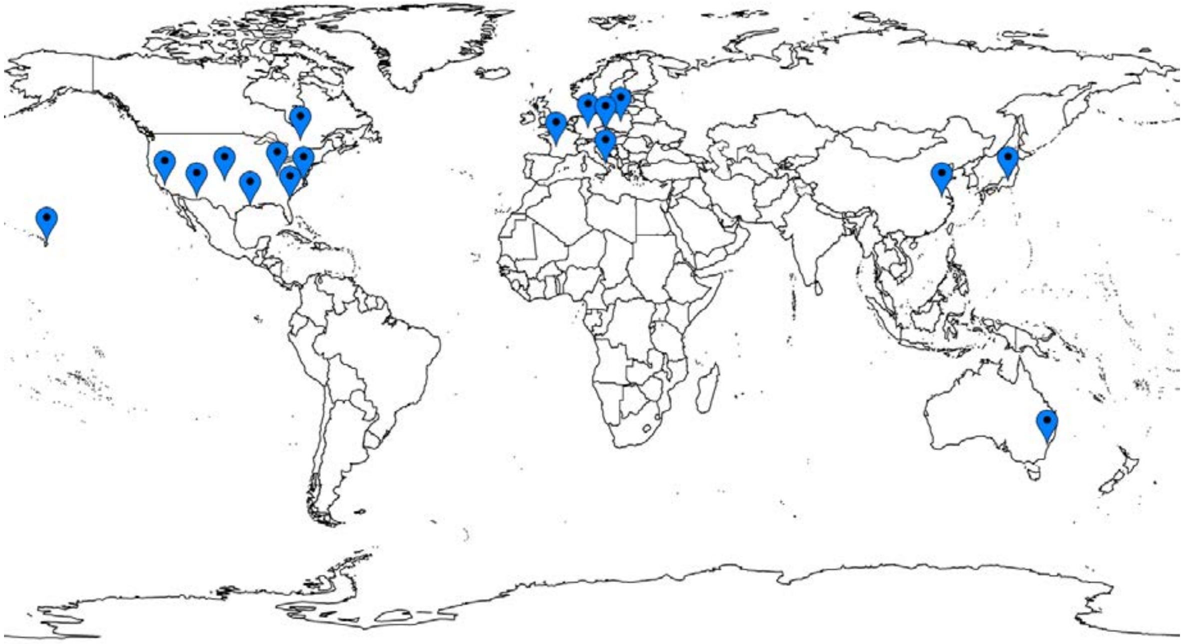
David Moroni/JPL
Kevin Murphy/NASA
Charles Norton/JPL
Jean-Francois Piolle/IFREMER
Nga Quach/JPL
Brandi Quam/NASA
Shawn Smith/FSU
Ben Smith/JPL
Adam Stallard/FSU

Rob Toaz/JPL
Vardis Tsontos/JPL
Suresh Vannan/JPL
Jorge Vazquez/JPL
Ou Wang/JPL
Brian Wilson/JPL
Steve Worley/NCAR
Elizabeth Yam/JPL
Phil Yang/GMU

# In Summary

- **You've got to think about big things while you're doing small things, so that all the small things go in the right direction** – Alvin Toffler
- Focus on end-to-end data and computation architecture, and the total cost of ownership
- JPL Strategy is to drive Data Science into the fabric of JPL by
  - Launching cross-institution pilots
  - Building a trained workforce
  - Linking to the mission-science data lifecycle
- Invest in Interactive Analytics that simplifies the integration of *multiple* Earth observing remote sensing instruments; comparison against models

# Presentation | Evaluation | Collaboration



Presentation | Evaluation | Collaboration

**Thomas Huang**

thomas.huang@jpl.nasa.gov

Jet Propulsion Laboratory

California Institute of Technology